Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 2

11-1-2010

The Not-So-Quiet Revolution: Cautionary Comments on the Rejection of Hypothesis Testing in Favor of a "Causal" Modeling Alternative

Daniel H. Robinson
University of Texas, jrlevin@u.arizona.edu

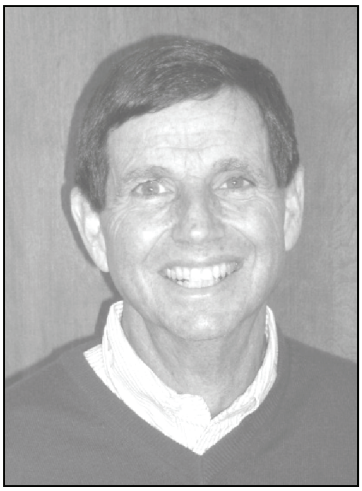
Joel R. Levin
University of Arizona, jrlevin@u.arizona.edu

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 332-339

1538 – 9472/10/$95.00

INVITED DEBATE

The Not-So-Quiet Revolution: Cautionary Comments on the Rejection of Hypothesis Testing in Favor of a "Causal" Modeling Alternative



Daniel H. Robinson
University of Texas

Joel R. Levin
University of Arizona

Rodgers (2010) recently applauded a revolution involving the increased use of statistical modeling techniques. It is argued that such use may have a downside, citing empirical evidence in educational psychology that modeling techniques are often applied in cross-sectional, correlational studies to produce unjustified causal conclusions and prescriptive statements.

Key words: Modeling, hypothesis testing, SEM, HLM, causation.

Daniel H. Robinson is Professor of Educational Psychology and editor of *Educational Psychology Review*. His research interests include educational technology innovations that may facilitate learning, team-based approaches to learning, and examining gender trends concerning authoring, reviewing, and editing articles published in various educational journals and societies. Email him at: dan.robinson@mail.utexas.edu.

Joel R. Levin is Professor Emeritus at the University of Arizona and the University of Wisconsin-Madison. He is former editor of the *Journal of Educational Psychology*, and former Chief Editorial Advisor for journal publications of the American Psychological Association. His research interests include the design and statistical analysis of educational research, as well as cognitive-instructional strategies that improve students' learning. Email him at: jrlevin@u.arizona.edu.

Introduction

Over the years, we have found that Joseph Rodgers (e.g., Rodgers, Cleveland, van den Oord, & Rowe, 2000; Rodgers & Nicewander, 1988) has something academically interesting, meaty, and instructive to say. Against that backdrop, Rodgers' most recent essay, provocatively titled "The epistemology of mathematical and statistical modeling: A quiet methodological revolution" (Rodgers, 2010) merits close examination and extensive

commentary. Rodgers appeared to have missed the mark in two critical respects; both reflected in the subtitle "A quiet methodological revolution," because as will become apparent in the following discussion, the revolution is neither quiet nor methodological.

The Null Hypothesis Hullabaloo

Rodgers is correct in stating that serious concerns about null hypothesis significance testing (NHST) have been mounting over the past several decades. Yet, as is well represented in Harlow, Mulaik, & Steiger's (1997) impressive volume, NHST criticisms have hardly been expressed quietly, but rather with full sound and fury. Moreover, in making his case, Rodgers provided a one-sided view of the controversy. Although several sources that indict NHST were cited, short shrift was given to approaches that have defended reasonable and proper applications of statistical hypothesis testing, including, among others, deciding whether a "believed-random" process is truly random (e.g., Abelson, 1997), "intelligent hypothesis testing" (Levin, 1998a), "equivalence testing" (e.g., Serlin & Lapsley, 1993), and hypothesis testing supplemented by effect-size estimation and/or confidence-interval construction (Steiger, 2004).

In addition, numerous authors have defended the use of NHST when mindfully applied (e.g., Frick, 1996; Hagen, 1997; Robinson & Levin, 1997; Wainer & Robinson, 2003). Rodgers cited social-sciences statistical sage Jacob Cohen (1994) as one who dismissed NHST practices in his 1994 seminal article, "The Earth is Round ($p < .05$)." Yet, in the same article, one could easily interpret Cohen's (p. 1001) comment about the "nonexistence of magical alternatives to NHST" as conceding that for whatever "good" NHST does, there are no adequate substitutes.

Rodgers (p. 2) described the fundamental difference between the Fisherian and Neyman-Pearson approaches, with the latter "emphasiz[ing] the importance of the individual decision." However, he characterized NHST as a hybrid and condemned it. Just because a technique is often misused is not a sufficient reason to abandon it. For example, it is argued below that in educational psychology we have observed frequent misapplication of the Rodgers' favored causal modeling techniques. In recognizing that misapplication, however, our goal is not to deter researchers from adopting modeling techniques, but rather to encourage researchers to apply such techniques appropriately and to interpret wisely the results that they pump out. (Back in the Neanderthal age of computers, "grind out" would have been a much more fitting description.)

As researchers who have spent most of our careers conducting randomized experiments, we have sought to apply NHST judiciously, typically adopting or adapting Neyman-Pearson *a priori* Type I, Type II error, effect-size, and sample-size specification principles. Accordingly, we have found that in experiments conducted with rationally (or better, optimally) determined sample sizes - that is, sample sizes associated with enough statistical power to detect nontrivial differences but with not too much power to detect trivial differences (see, for example, Levin, 1998b; and Walster & Cleary, 1970) - NHST provides useful information concerning whether one has an experimental effect worth pursuing. In this context, pursuing means that obtaining a statistically significant effect is followed by a sufficient number of independent replications until the researcher has confidence that the initially observed effect is a statistically reliable one (see, for example, Levin & Robinson, 2003).

In that sense as well, we have regarded NHST primarily as a screening device, similar in function to what Sir Ronald had in mind (e.g., Fisher, 1935). Much of the hullabaloo about NHST is caused by too many researchers focusing on the results of a single study rather than on a series of studies that are part of a program of research (Levin & Robinson, 2000). Fisher was never satisfied with an effect identified in a single study, even if it had a p value of less than 0.05! Instead, he believed that a treatment was only worth writing home about when it had consistently appeared in numerous experiments. As is implied in the following section, whatever purported advantages modeling techniques have over NHST also vanish unless researchers test *a priori* models in multiple experiments.

Rodgers (p. 3) also condemned the

NHST jurisprudence model while aptly referring to Tukey's (1977) "confirmatory data analysis" strategy as being judicial (or quasi-judicial) in nature. Yet, Rodgers mischaracterized Tukey's exploratory data analysis strategy insofar as the detective nature of that hypothesis-generating approach clearly is not jurisprudence. It is this detective role that one emphasizes when using NHST simply as a research-based screening process to determine whether posited effects exist. To us, convincing a jury of one's peers that a prescription for practice should be based on a single research study is rarely, if ever, justified.

Rodgers' (p. 9) assertion that a fundamental problem with NHST is one of testing valueless nil null hypotheses has been advanced by many critics. As researchers who endeavor to use intelligent forms of hypothesis testing with experimental data, we regard the problem of nil nulls not as a statistical issue but as a methodological one. Specifically, it makes little or no conceptual sense to apply NHST when comparing an instructional treatment with a "closet" (Levin, 1994, p. 233) control group (i.e., a condition in which participants sit in a dark room and do nothing), just as it is inane to compute *p*-values for reliability correlations (see, for example, Thompson, 1996). Educational psychology is filled with such examples of comparing new innovations with ridiculous straw-person control conditions that no sane researcher would ever consider using. A more appropriate formulation of a nil null is when an investigator wishes to compare a newly developed and previously untested experimental treatment with the best treatment that is currently available.

According to Rodgers, "the [1999 task force assembled by the American Psychological Association] concluded that NHST was broken in [a] certain respect" (p. 3). Task-force member Wainer and the present first author (Wainer & Robinson, 2003) provided a different view of the task force's brief consideration of the recommendation to issue an outright ban on NHST. As we have argued previously (e.g., Levin & Robinson, 1999) and in our preceding discussion, adopting such an extreme stance would be akin to calling for a ban on hammers because hammerers were hammering their fingers instead of nails (for additional discussion, see Levin & Robinson, 2003). Even the outspoken NHST critic Rozeboom (1997) acknowledged via another "tools" analogy that "the sharpest of scalpels can only create a mess if misdirected by the hand that drives it," (p. 335). Fortunately, in the case of the most recent (6th) edition of the *APA Publication Manual* (American Psychological Association, 2010), the hypothesis-testing baby was not thrown out with the bath water.

"Causal" Modeling Techniques

Contemporary modeling techniques, including structural equations modeling (SEM) and hierarchical linear modeling (HLM), among others, which emerge from a theoretical/conceptual framework, are statistical/data-analytic and not methodological in nature. So, whence Rodgers' "methodological" revolution? Even he noted on p. 8 that "SEM has been built into a powerful analytic method and is a prototype of the first approach [a model-comparison framework] to postrevolutionary modeling" (p. 8).

That a statistical modeling tail often wags the methodological dog may have contributed to what we consider a major misuse of causal modeling: researchers attempting to squeeze causality out of observational or correlational data. Because of the unfortunate "causal" nomenclature, we fear that many researchers may be deluded into believing that the statistical control that such techniques provide for correlational (non-experimental) data is on a par with the genuine experimental control of randomized experiments (Levin & O'Donnell, 2000, p. 211). This in turn results in causal-model appliers issuing causal conclusions that they mistakenly believe are scientifically valid. As Cliff (1983) previously noted, "Literal acceptance of the results of fitting 'causal' models to correlational data can lead to conclusions that are of questionable value" (p. 115).

In addition, because causal-model researchers' conclusions typically flow from revised data-driven models rather than from *a priori* theory-based model specifications, in the absence of independent validations those causal conclusions present even more cause for

concern. As with our previous hammers vs. hammerers distinction, Rodgers is well aware of researchers' potential shoddy application of causal modeling techniques. Yet, he could have sent a stronger cautionary message to the relatively uninitiated model builder than his innocuous pronouncement that "the success of SEM depends on the extent to which it is applied in many research settings" (p. 8).

To illustrate what we mean by prescriptive statements appearing in articles that include statistical modeling techniques, we offer very recent examples that appeared in a reputable educational psychology research journal. To avoid redundancy, we offer only two such unjustified causal excerpts here, from numerous ones that we have encountered in multiple teaching-and-learning research journals that we have recently read or reviewed (see Robinson, Levin, Thomas, Pituch, & Vaughn, 2007, and the following section).

Ciani, Middleton, Summers and Sheldon (2010)'s Study

The following summary appeared in Ciani et al.'s study abstract:

> Multilevel modeling was used to test student perceptions of three contextual buffers: classroom community, teacher's autonomy support, and a mastery classroom goal structure…Results provide practitioners with tools for counteracting potential negative implications of emphasizing performance in the classroom. (p. 88)

There was one predictor variable; one outcome variable, a three-item scale that measured students' motivation to learn; and three moderator variables, a three-item scale that measured student perceptions of classroom community, a four-item scale that measured student perceptions of instructor autonomy support, and a three-item scale that measured student perceptions of the extent to which their teacher emphasizes developing competence in the classroom. All measures were collected at a single point in time and HLM was used to analyze the data. Here are a couple causal conclusions from the discussion section:

> However, it appears that comparing students' achievement publicly, or using the work of the highest achieving students as an example for everyone, may not be so pernicious a practice when students in the classroom perceive a sense of community among their fellow classmates.
>
> [O]ur findings demonstrate that if students feel respected by the teacher, such that their preferences and ways of doing things are acknowledged and accommodated as much as possible, then a strong performance orientation on the part of the teacher is not harmful. Autonomy support enables students to internalize what they are doing, so that they view their activity as important even if it is not enjoyable, or if it creates stress and pressure. Thus, it appears that emphasizing competition between students is not necessarily undermining of student mastery goals, if the teacher can communicate and promote the performance structure in a non-controlling way. These findings are reassuring, showing that performance orientations are not necessarily corrosive – certainly an important message, given the performance necessities that all students face. (p. 95)

As with most of these articles based on correlational data and yet that offer prescriptive recommendations, certain limitations of the research are explicitly acknowledged by the authors:

> The most significant limitation to the current study is that all data reported are correlational.
>
> Gathering data at one point in time also creates a limitation regarding the causal relationships among the variables in this study. (p. 96)

These limitations aside (or ignored?), the authors proceeded to offer the following prescriptive:

> Our findings, along with other goal theorists (e.g., Urdan & Midgley, 2003), suggest that given current prevailing attitudes and policy it may be more fruitful to emphasize adaptive instructional practices in the classroom, as opposed to trying to reduce maladaptive practices. (p. 97)

Thus, the authors made recommendations for practice ("prescriptive statements") in the absence of convincing evidence that such practices are clearly causally related to student outcomes.

Chen, Wu, Kee, Lin & Shui's (2009) Study

Chen et al. used SEM to analyze relations among fear of failure, achievement goals, and self-handicapping. Causal relations among the variables are implied in the Discussion section:

> This finding shows fear of failure as a distal determinant of self-handicapping and achievement goals (MAv and PAv) as proximal determinants of self-handicapping, demonstrating the motivational process of self-handicapping. (p. 302)

The authors revealed the perceived magical quality of SEM allowing researchers to coax causality from correlational data:

Since SEM analysis examines many variables' relationships simultaneously, we rely on its results as the basis for our conclusions and discussion. (p. 303)

The Limitations section is predictable:

> Although we used the SEM approach to estimate the proposed model, the data in the study are cross-sectional in nature and causal relations cannot be drawn. The longitudinal approach is preferred in order to ascertain the causal pattern and to further clarify the chronic effects of mastery-avoidance and performance-approach goals on achievement-related outcomes. (p. 304)

In contrast, what follows are the grand prescriptives that appeared in the Implications and Conclusions:

> We believe that the integrative model can help educators develop effective interventions to reduce students' self-handicapping, especially since we found that the mid-level achievement goals (MAv and PAv) mediate the relationships between fear of failure and self-handicapping… it is suggested that teachers use multiple indices to offer more opportunities for students to attain success. In addition, teachers should encourage students to embrace a multiple goals perspective in which doing one's best and outperforming others are not in conflict with each other. (p. 304)

Rodgers (2010, p. 8) previously proffered caveat aside, in both of the just-presented examples, cross-sectional (one time point), correlational (no variables were manipulated) data were tossed into a statistical modeling analysis and what popped out were causal conclusions.

Correlational Data and Causal Conclusions

Over the past few years, we have examined empirical articles published in widely read teaching- and-learning research journals and have found that:

1. In one journal survey (Hsieh et al., 2005), the proportion of articles based on intervention and experimental (random assignment) methodology had decreased from 47% in 1983 to 23% in 2004.

2. In another journal survey (Robinson et al., 2007), the proportion of articles based on intervention methods had decreased from 45% in 1994 to 33% in 2004. Meanwhile, the proportion of nonintervention articles that contained prescriptive statements increased from 34% in 1994 to 43% in 2004. The proportion of nonintervention (non-experimental and correlational) articles that included prescriptive statements (in the form of causally implied implications for

educational practice) increased from 33% in 1994 to 45% in 2004.

3. In a follow-up to the just-described Robinson et al. (2007) survey (Shaw, Walls, Dacy, Levin & Robinson, 2010), although only 19 nonintervention studies in 1994 included prescriptive statements, these statements were repeated in 30 subsequent articles that had cited the original 19.

For the present article, we examined the first two issues of the 1999 volume of the APA-published journal, the *Journal of Educational Psychology*, and again for the 2009 volume. We looked specifically at the comparative proportions of articles based on correlational methods and those that involved interventions (either randomized experimental or nonrandomized but researcher manipulated), as well as the proportion of correlational methods articles in which prescriptive statements were offered. The results are summarized in Table 1.

Although roughly half of the articles appearing in only one of the five journals that were part of Robinson et al.'s (2007) study were surveyed, the findings support the reported trends. Intervention studies (both randomized and nonrandomized) are becoming increasingly rare and instead researchers are basing their recommendations for practice on weaker evidence. Moreover, it appears that statistical and nonrandomized) are becoming increasingly rare and instead researchers are nonrandomized) modeling techniques are becoming more popular - having increased from only 3% of the correlational research articles in 1999 to 40% in 2009 - which may in turn contribute to the concomitant 10-year increase in prescriptive statements appearing in such articles.

Thus, we have witnessed widespread application of SEM, HLM, and other sophisticated statistical procedures in correlational data contexts, where causality is sought but the critical conditions needed to attribute causality are missing (e.g., Marley & Levin, 2011; Robinson, 2010). Rodgers states that "researchers who are scientists…should be focusing on building a model…embedded within well-developed theory" (p. 4-5). Here we agree with former Institute for Educational Science Director Grover Whitehurst who argued that - at least in the field of education - we have enough theory development studies and need more studies that address practical "what works" questions.

It is our fear that a research approach where the question, "Does the data fit my model?" is far more dangerous than the question, "Is there anything here worth pursuing?" As we have seen, an affirmative answer to the former question seems to entitle a researcher to form a model that indicates a causal relationship between, say, students' self-efficacy and their achievement. The researcher then develops a self-efficacy scale that measures

Table 1: Summary of Selected Results of Surveyed Articles Appearing in the *Journal of Educational Psychology* (1999 and 2009) Based on Either Correlational or Intervention Methods

	1999		2009	
	Type of Study		Type of Study	
	Correlational	Intervention	Correlational	Intervention
Number of Articles	18 (60%)	12 (40%)	23 (66%)	12 (34%)
Prescriptive Statements	9 (50%)	------[a]	13 (57%)	------[a]
Statistical Modeling	1 (3%)	0 (0%)	14 (40%)	2 (6%)
Prescriptive Statements	1	------[a]	7 (50%)	------[a]

Note: This table includes preliminary data from a larger study recently completed by Reinhart, Haring, Levin, Patall, and Robinson (2011). [a] Not assessed in the present survey

students' self-perceptions and also measures achievement. The data may fit the model but in the absence of convincing longitudinal data, ruling out alternative explanations, and independent replications based on the previous nice-fitting model, this practice may lead to dangerous causal conclusions. For the just-presented self-efficacy example, it is just as likely that high achievers feel better about their effectiveness as learners rather than the other way around. Apparently, many researchers believe that it is entirely appropriate to apply such modeling techniques and to interpret the results as support for prescriptive statements founded on causality.

Conclusions About Revolutions

To summarize, Rodgers (2010) has written a cogent essay on the vices of statistical hypothesis testing and the virtues of statistical modeling. We believe, however, that his essay painted a somewhat distorted (and potentially misleading) portrait about those statistical "arts." In particular, we take issue with two aspects of Rodgers' so-called "quiet methodological revolution." For one aspect (rejecting statistical hypothesis testing), we argue that the picture is neither as bleak nor as open and shut as Rodgers portrayed. As supporting evidence, witness the sustained presence of hypothesis testing, along with its more intelligent additions and adaptations, in various academic-research disciplines - including the research-and-publication "bible" of both our very own field of psychology and virtually all social-sciences domains, the most recent edition of the APA *Publication Manual* (American Psychological Association, 2010).

For the other aspect of Rodgers' essay that merits critical commentary (accepting modeling techniques), we argue that causal modeling and other related multivariate and multilevel data-analysis tools frequently cause their users to think - in accord with Rodgers' seductive subtitle - that the procedures are methodological randomization-compensating panaceas rather than techniques that do the best they can to provide some degree of statistical control in a "multiply confounded variable" world. The unfortunate consequence of that methodological understanding, then, is that when combined with researcher misapplication of such modern modeling artillery, instead of being on target with their data analyses and research conclusions, weapons are backfiring and researchers are ending up (whether knowingly or not) with a considerable amount of egg on their faces.

References

Abelson, R. P. (1997). The surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (*6th Ed.*). Washington, DC: American Psychological Association.

Chen, L. H., Wu, C.-H., Kee, Y. H., Lin, M.-S., & Shui, S.-H. (2009). Fear of failure, 2 x 2 achievement goal and self-handicapping: An examination of the hierarchical model of achievement motivation in physical education. *Contemporary Educational Psychology*, *34*, 298-305.

Ciani, K. D., Middleton, M. J., Summers, J. J., & Sheldon, K. M. (2010). Buffering against performance classroom goal structures: The importance of autonomy support and classroom community. *Contemporary Educational Psychology*, *35*, 88-99.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Cohen, J. (1994). The earth is round ($p <$ .05). *American Psychologist*, *49*, 997-1003.

Fisher, R. A. (1935). *The design of experiments*. (Reprinted in 1960). Edinburgh: Oliver & Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test, *American Psychologist*, *52*, 15-24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Hsieh, P., Hsieh, Y. P., Chung, W. H., Acee, T., Thomas, G. D., Kim, H. J., You, J., Levin, J. R., & Robinson, D. H. (2005). Is educational intervention research on the decline? *Journal of Educational Psychology*, *97*, 523-529.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *5*(*2*), 43-53.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gaps? *Issues in Education: Contributions from Educational Psychology*, *5*, 177-229.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*, *11*, 143-155.

Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, *29*(1), 34-36.

Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods*, *2*, 231-236.

Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review*, *23*, 197-206.

Reinhart, A. L., Haring, S. H., Levin, J. R., Patall, E. A., & Robinson, D. H. (2011). *Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data*. Unpublished manuscript, University of Texas, Austin.

Robinson, D. H. (2010, May). *Correlational, causal, and prescriptive claims: Guidelines for articles appearing in* Educational Psychology Review. Paper presented at the annual meeting of the American Educational Research Association, Denver.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*, 21-26.

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. R. (2007). The incidence of "causal" statements in teaching and learning research journals. *American Educational Research Journal*, *44*, 400-413.

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, *55*, 599-612.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*, 59-66.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*, 335-391. Mahwah, NJ: Erlbaum.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues*, 199-228. Hillsdale, NJ: Lawrence Erlbaum.

Shaw, S. M., Walls, S. M., Dacy, B. S., Levin, J. R., & Robinson, D. H. (2010). A follow-up note on prescriptive statements in nonintervention research studies. *Journal of Educational Psychology*, *102*, 982-988.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164-182.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(*2*), 26-30.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, *32*(*7*), 23-31.

Walster, G. W., & Cleary, T. A. (1970). Statistical significance as a decision-making rule. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology*, 246-254. San Francisco: Jossey-Bass.